Network inference from dynamic (state) information

Read Chapter 13 of our textbook

Input: components; states of components (in time)

Hypotheses: regulatory framework

Output: proposed regulatory network

Validation: capture known interactions

For inference of gene regulatory networks, the most frequently used state information comes from gene expression arrays (microarrays)

There are several microarray types and methods, for our purposes it suffices to say that a microarray provides a readout of the relative or (semi)absolute expression level of each gene in the array.

Analysis of differential gene expression is not network inference!

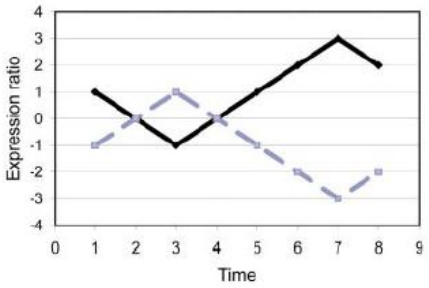
Inference methods

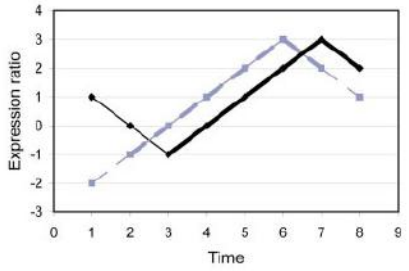
- Need expression snapshots:
  - Correlation analysis
  - Bayesian networks
- Need expression timecourse:
  - Continuous – Differential equations
  - Discrete - Boolean
- Need other types of information:
  - Data mining

The problem is under-constrained regardless of the choice of methods: the number of conditions or timepoints is less than the number of degrees of freedom in the system. Thus it's usually impossible to find a unique solution.

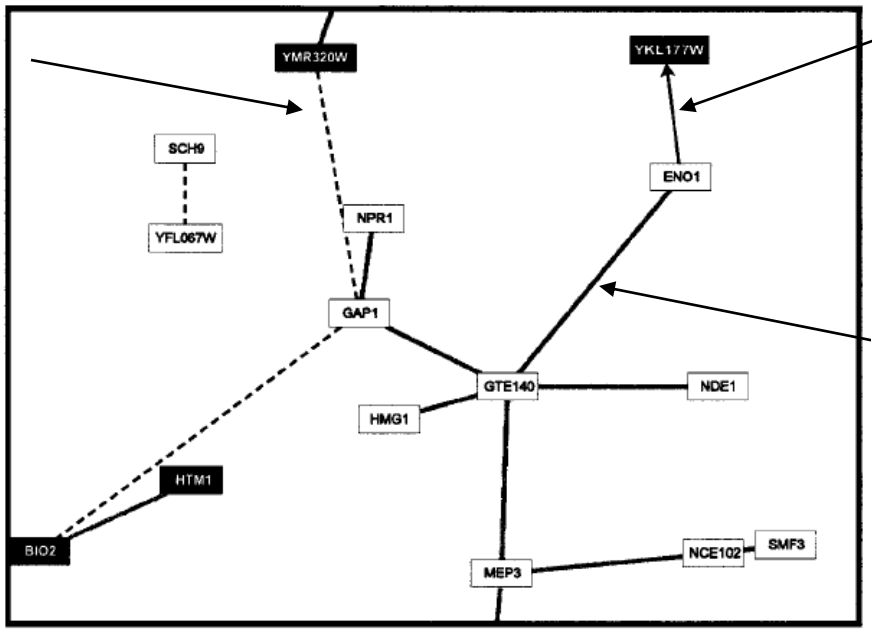
Correlation (co-expression) analysis

- Pairwise correlation of expression levels of two genes across time or conditions, e.g. by Pearson correlation or Euclidean distance
- Negative correlations or time-delayed correlations are also informative





Inverted relationships
in the expression profiles.



Time delayed
expression

Simultaneous
expression

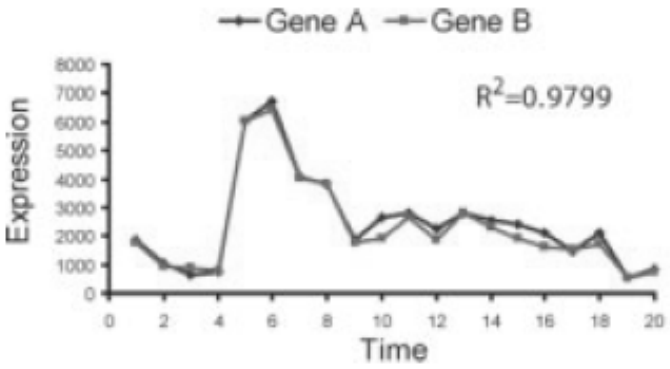
Qian *et al*. (2001) J Mol. Bio 314, 1053-1066

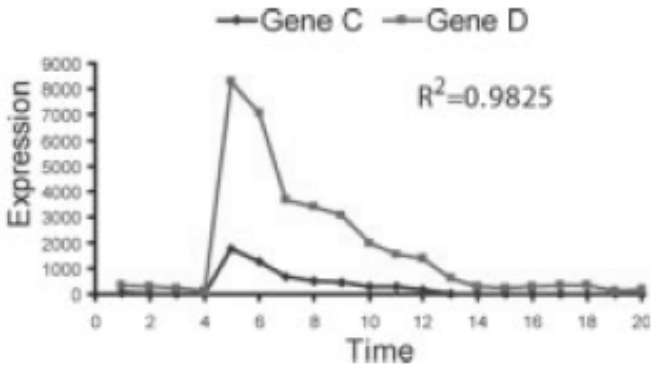
Drawback: little causal insight.
Co-expressed genes may not be co-regulated.

Estimation of directionality

Measure of correlation: $R^2 = b_{YX} * b_{XY}$
b_{YX}, b_{XY} *:* regression slopes (regressing Y on X and X on Y)

R^2 values cannot differentiate between expression levels





Slope ratio metric: divide the smaller slope by the larger

e.g. b_{YX} = 1.004 *and* b_{XY} = *0.976* SR_{XY} = 0.97

Directionality is assigned to those edges for which $SR \to 0$
Direction assumed to be from gene with lower expression to the one with higher.
(Can you guess the justification?)

Gupta, A. *et al*. (2006) Bioinformatics 22, 209-214

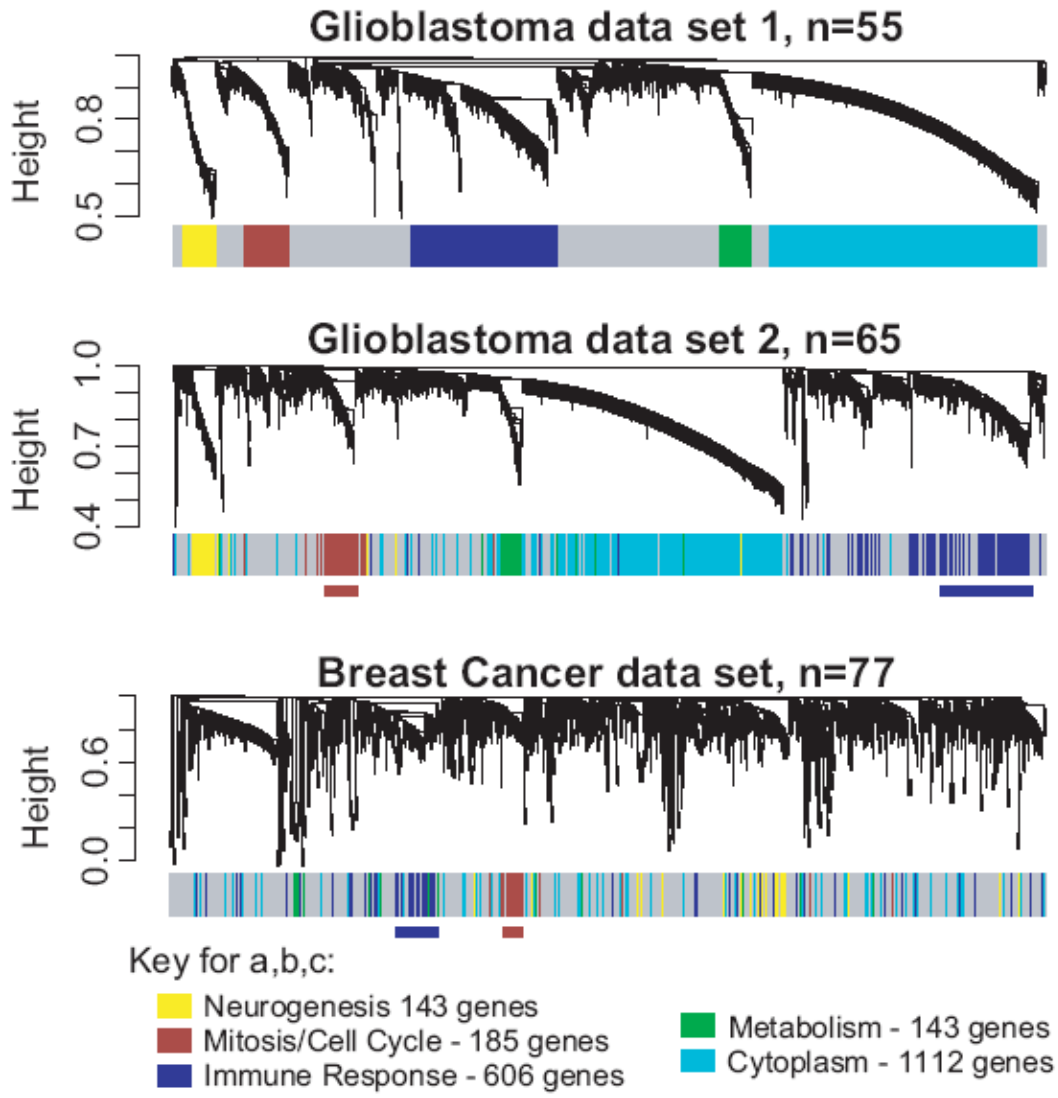
Clustering analysis

- The pairwise correlation among genes across time or conditions can be used as basis for clustering algorithms. The hope is that clusters will correspond to functional modules.

- Hierarchical clustering - forms a dendogram
  - Successive clusters are formed by aggregation of existing clusters.
  - Difficult to decide which level in the dendogram is the best

- K- means clustering
  - K - predetermined number of groups
  - Clusters should be internally similar but externally dissimilar.
  - Start with random assignment, iteratively refine.
  - More computationally intensive than hierarchical clustering but optimization can be performed.

Oncogenic clustering analysis



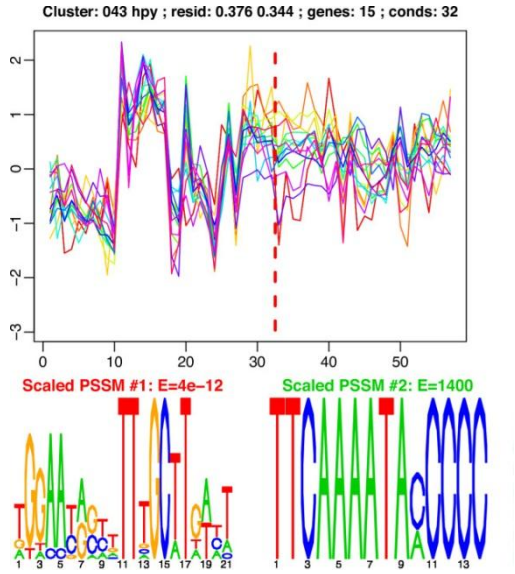
- Constructed weighted gene co-expression network based on pairwise Pearson correlations
- Assigned thresholds such that the network is scale free.
- Hierarchical clustering to detect groups of co-expressed genes.
- The five co-expressed groups in data set 1 map relatively well to the other data sets.

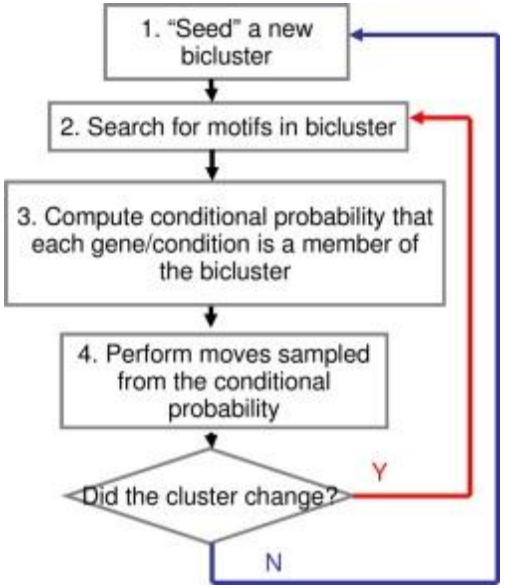
Horvath, S. *et al*. (2006) 103, 17402-7

Bi-clustering

- Cluster both genes and conditions at the same time
- Genes can be part of multiple clusters
- Example: cMonkey combines expression information with shared transcriptional motifs

- The detected clusters can be considered as single nodes, and other methods can be used to infer edges among them.



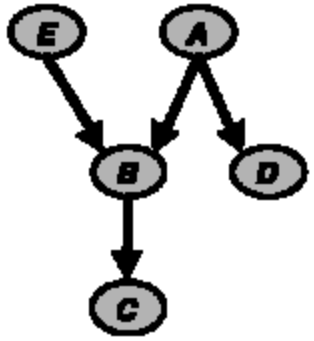


Reiss et al BMC Bioinf. 2006

Bayesian networks

- Probabilistic approach based on *dependence* and *conditional independence* in the data
- Estimates the confidence in the different features of the network

- Main step: construct a directed acyclic graph indicating dependencies.



- A scoring function is designed to evaluate each candidate network with respect to the training data and search for the optimal network.

$$P(A, B, C, D, E) = P(A)P(B|A, E)P(C|B)P(D|A)P(E).$$

- Start with random or heuristic graph, change edges iteratively. The graph yielding the highest Bayesian score is chosen as the best fit to the data.

Deterministic Methods for Network Inference

Deterministic inference correlates the rate of change in expression level of each gene with the levels of other genes by finding the functional or logical forms of these interdependence relationships.

Can only be applied if time-course expression data is available.

(Loosely) Two classes of deterministic inference methods:

1) **Continuous;**
2) **Discrete**

Continuous Methods

- Systems of linear or nonlinear differential equations in which the rate of change of expression of X_i(t) is a combination, (e.g., linear), of concentrations of all other X(t) :

$$\frac{dX_i(t)}{dt} = \sum_{j=1}^{N} w_{ji} X_j(t)$$

- **Pros and cons**:

  -accuracy increases as number of experimental time points increases;

  -computational intractability quickly becomes an issue

Example: Inferring gene-regulatory networks in *B. subtilis* using a linear model

Microarray data

Microarray data

$$\frac{dZ_i(t)}{dt} = \sum_{j=1}^{N} w_{ji} Z_j(t)$$

To be solved for regulatory coefficient

$w_{ji} > 0 \Rightarrow$ ***activation of i by j***

$w_{ji} < 0 \Rightarrow$ ***inhibition of i by j***

Gupta, A., Varner, J. D., and Maranas, C. D, Computers and Chemical Engineering, 2005.

Network filtering

Multiple solutions of this equation: all possible alternate network configurations that are consistent with the experimental data.
Goal: find the sparsest network

Used linear programming (LP)

- To get sparse network we want to maximize number of zero weights
- The objective function minimizes deviation of weights from zero

$$\underset{c_{jk},w_{ij}^{+},w_{ij}^{-}}{\text{minimize}} \quad \sum_{i,j}(w_{ij}^{+}+w_{ij}^{-})$$

The model is able to identify hub regulators and interactions of highly expressed genes, e.g. genes involved in information processing, energy metabolism and signal transduction

Inferelator

- Use cMonkey to find gene bi-clusters

$$\frac{dy}{dt} = -y + \sum_{j=1}^{n} \beta_j X_j$$

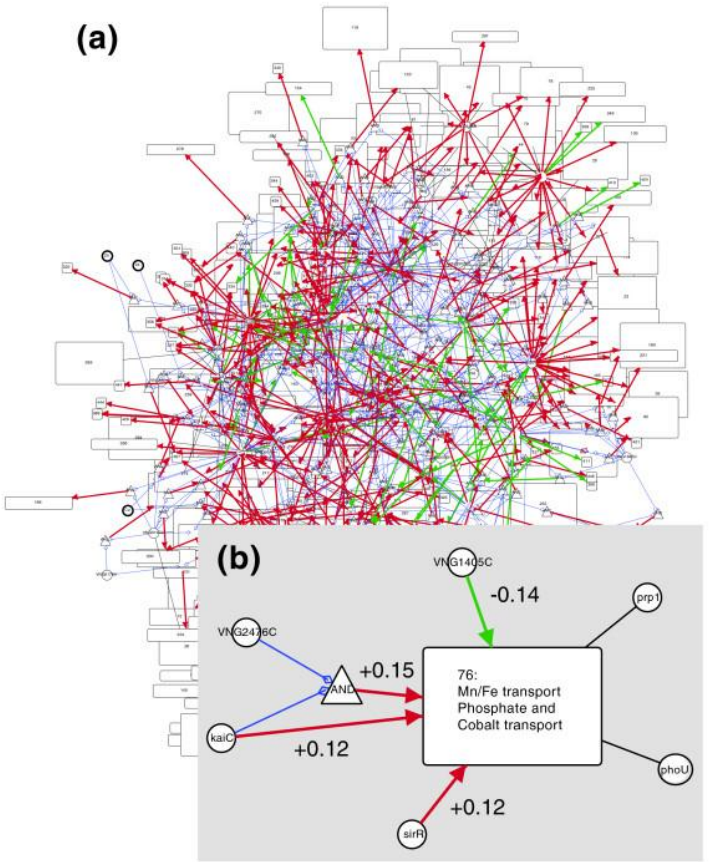
Mean expression level of a group of co-regulated genes

Regulatory factors

- Can also use more complex, nonlinear functional forms
- Used this method to infer a Halobacterium NRC-1 network

Circles: regulators, boxes: bi-clusters

Bonneau et al., Genome Biology 2006



Discrete Methods

- Boolean and other logic-based methods that predict discrete regulatory relationships

  - **Pros and cons:**

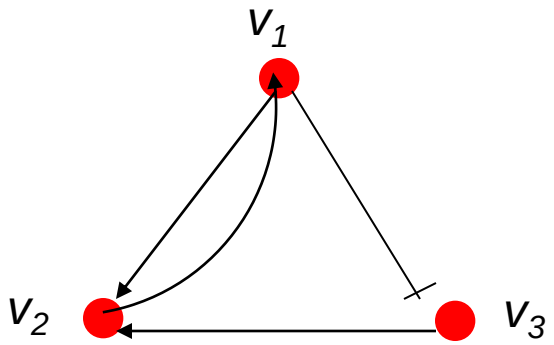
    -More computationally-tractable than continuous methods;

    -Less accurate than continuous methods.

  - **Implemented algorithms for large-scale inference:** e.g. REVEAL (REVerse Engineering ALgorithm)

Liang, S., Fuhrman, S., and Somogyi, R, Pac Symp Biocomput, 1998, pp. 18-29

Example of Boolean model



RULES

$$v_1^* = v_2, \; v_2^* = v_1 \; AND \; v_3, \quad v_3^* = NOT \; v_1$$

	INPUT			OUTPUT	
v_1	v_2	v_3	v_1*	v_2*	v_3*
0	0	0	0	0	1
0	0	1	0	0	1
0	1	0	1	0	1

Example of Boolean inference

	Genes at time *t*			**Genes at time *t+1***			
	v_1	v_2	v_3	v_1^*	v_2^*	v_3^*	
I_1	1	0	0	0	0	1	O_1
I_2	0	1	0	0	1	1	O_2
I_3	0	1	1	1	0	0	O_3

G_1

$$v_1^* = v_3$$
$$v_2^* = v_2 \ AND \ (NOT\ v_3)$$
$$v_3^* = NOT\ v_3$$

Consistent

G_2

$$v_1^* = v_3$$
$$v_2^* = v_2$$
$$v_3^* = v_1 \ OR \ v_3$$

Not consistent

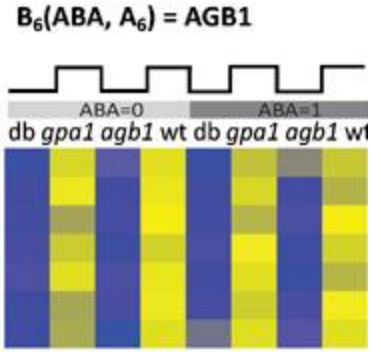
Conduct an exhaustive search through all Boolean rules for the nodes. Start with one input rules, then go to two input rules, until a cutoff. Stop when a consistent set is found.

Different methods for deciding consistency: mutual information between input and output, minimum description length, best fit extension. Implementation in Matlab toolbox.

Inferring G-protein action on the transcriptome

Data: gene expression in four *Arabidopsis* genotypes (Col, *gpa1*, *agb1*, *agb1 gpa1* double mutant), two tissues (guard cells and leaves) and two treatments (control and 50 μM ABA)

Hypotheses: gene expression patterns delineate genes (co)regulated by the G-protein and/or ABA



Inferred: putative G-protein and/or ABA regulatory modes and signaling pathways.

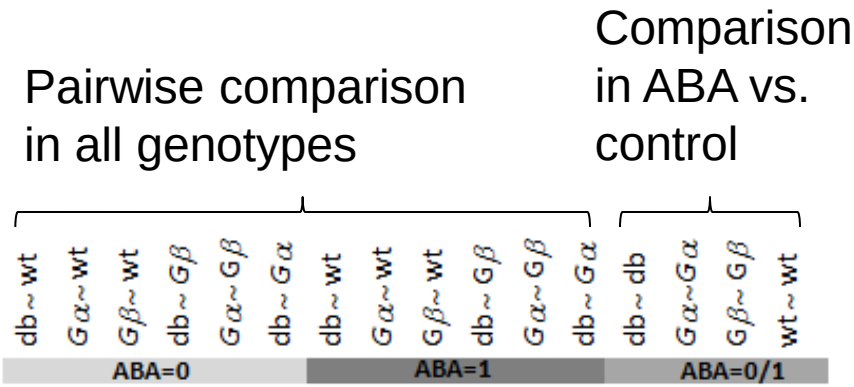
Validation: supports classical G-protein regulatory mechanisms.

Insight: identified a novel G-protein regulatory mechanism crosstalk between ABA and the signal that activates the G-protein system specificity in G-protein signaling

S. Pandey, R.S. Wang, L. Wilson, S, Li, T. E. Gookin, S.M. Assmann and R. Albert, Mol. Syst. Biology 6, 372 (2010).

Core of method: differential gene expression pattern

db= *gpa1 agb1*
Gα= gpa1
Gβ=agb1
wt = wild type



over-expressed

ABA-only regulation

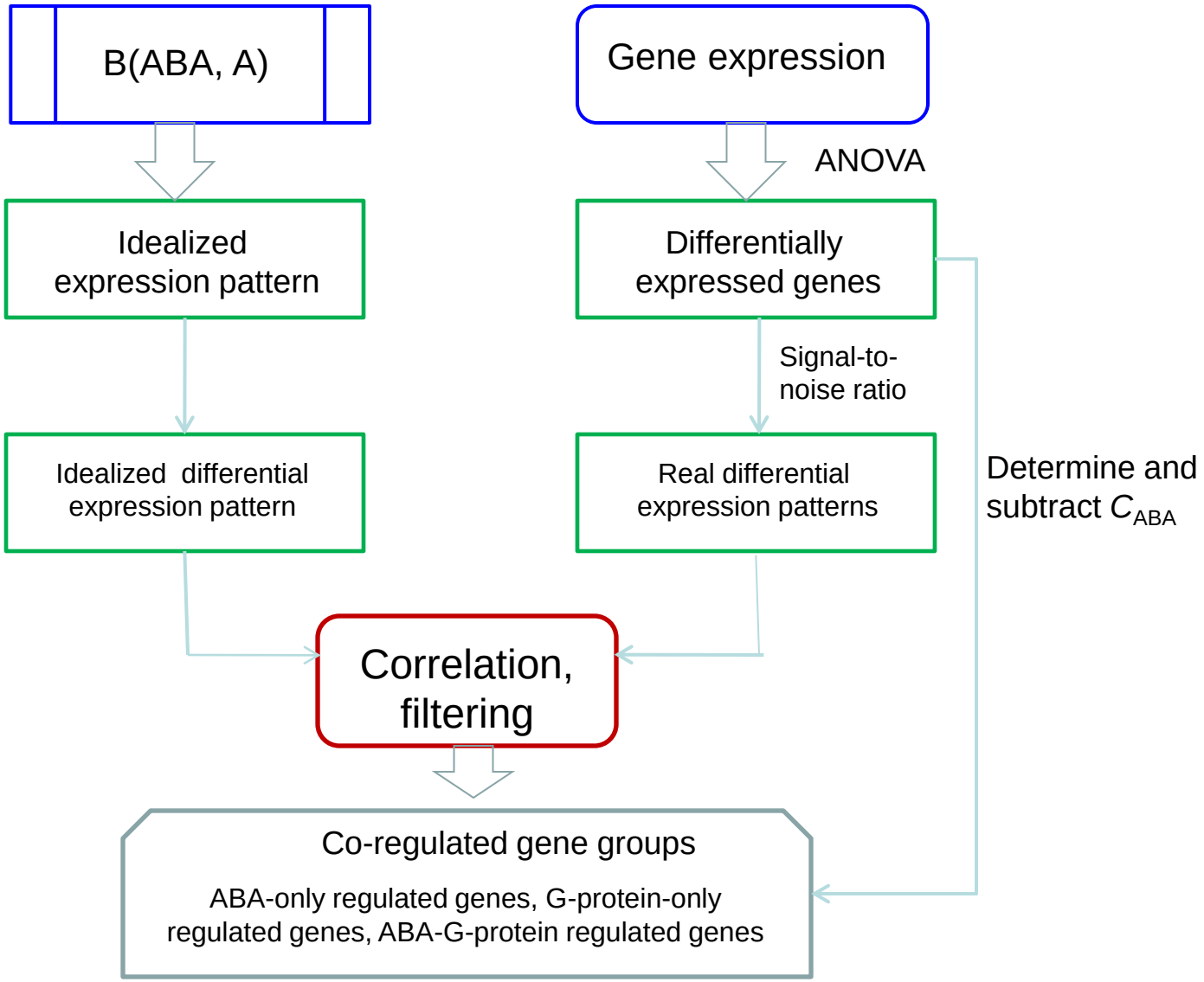
G protein-only regulation
Upregulation in *gpa1*

over-expressed

under-expressed

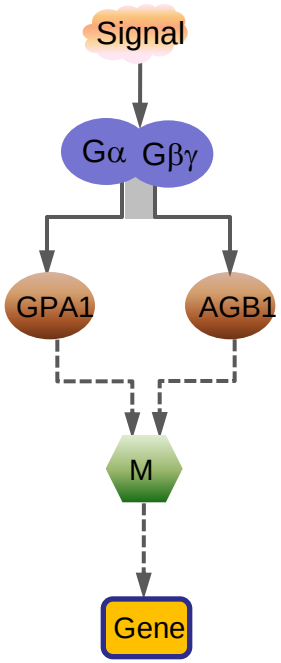
Mixed regulation

Overview of the method



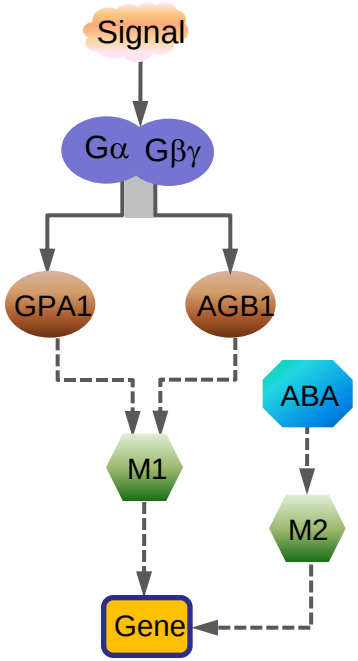
Relating co-regulated genes to signaling pathways

G protein regulation independent of ABA



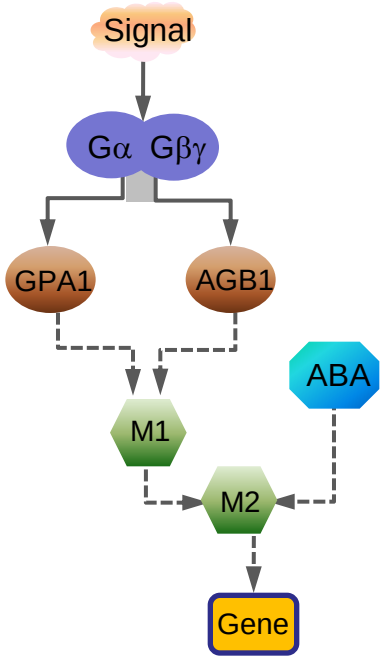
24 genes in guard cells and only 1 gene in leaves.
Not supported in leaves.

Additive effect of G protein and ABA



5 genes in guard cells and no genes in leaves.

Combinatorial regulation



Supported by 70 genes in guard cells and 470 genes in leaves.

Evidence of cross-talk between the signal activating the G protein and ABA.

Hybrid Methods

- Inference methods that bridge the gap between probabilistic and deterministic approaches, usually by incorporating some type of stochastic process (variability, uncertainty) into the inference algorithm.

- For example: Probabilistic Boolean Inference

- Pros and cons:
  - Arguably most accurate and realistic network inference methods;
  - Amount of training data and computational time make methods prohibitive for large networks;

Probabilistic Boolean Networks

- *N* Boolean functions are assigned to each node, each with some probability of being selected to advance the state of the node.
- The joint probability distribution of all Boolean functions corresponding to all nodes for the next time step can be calculated based on the present time step.

- A machine-learning algorithm must be used to update the state of each node at each time point. The main concept of the algorithm is the coefficient of determination that measures the extent to which a model is predictive of the value of the output.

Shmulevich, et al. (2002) Bioinformatics, 18(2): 261-274.

- As we have seen, probabilistic Boolean network are essentially the same as dynamic Bayesian networks.
- A PBN method was used to successfully infer a network regulating muscle development in Drosophila

Zhao, Serpedin, Dougherty (2006) Bioinformatics, 22: 2139.

Data mining

- Data-mining can be used to infer protein-protein interactions, gene-regulatory relationships, and even metabolic pathways.
- Extract information based on the statistical co-occurrence of features of interest e.g. their inclusion in databases and biomedical journals .
- In this case, by correlating the frequencies of keywords with the probability that a given interaction is addressed in a paper (estimated from a training set), machine learning algorithms can determine whether or not a particular paper is likely to discuss a specific interaction.
- Search tools such as STRING (http://string.embl.de/) employ similar data-mining methods for the inference of both direct and indirect protein-protein interactions in eukaryotes and prokaryotes.
- Example - Algorithm searched for the co-occurrence of pair of genes resulting in the edge generation according to the user defined threshold.

Marcotte, et.al. (2001) Bioinformatics, 17(4), 359-363

Other methods of network inference

- Assembly of causal but indirect effects into a sparse network
- Combination of pathway information with state data to infer best network and Boolean model

Inspiration: a model of drought signaling in plants

Phenomenon: abscisic acid induced closure of plant stomata

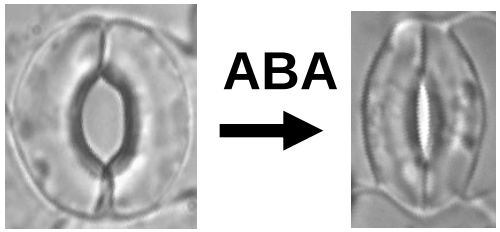
Hypotheses: network inference from indirect information

protein activity is switch-like (Boolean)

Validation: reproduces known wild type and disrupted behavior.

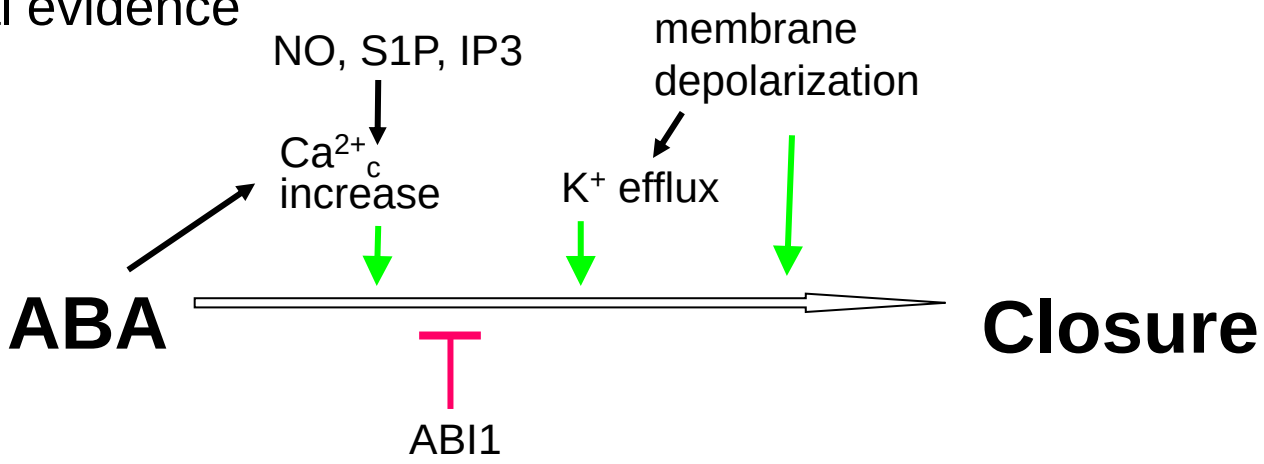
Explored: changes in initial conditions

changes in timing, disruptions



S. Li, S. Assmann and R. Albert, PLoS Biology 4, e312 (2006).

Literature information came from disruption experiments and leads to indirect causal evidence



Network construction from indirect evidence

- nodes: all proteins, molecules, ion channels implicated in the process
- compress biological information into activation or inhibition

Node A	**interaction**	**Node/Process B**	**species**
ABA	promotes	SphK	*Arabidopsis*
PLC	promotes	ABA → closure	*Commelina communis*
SphK	partially promotes	ABA → AnionEM	*Arabidopsis*

- hypothesis: indirect causal relationships and processes correspond to paths ABA ⟶⟶ ion flow, ABA ⟶⟶ Sph kinase activity

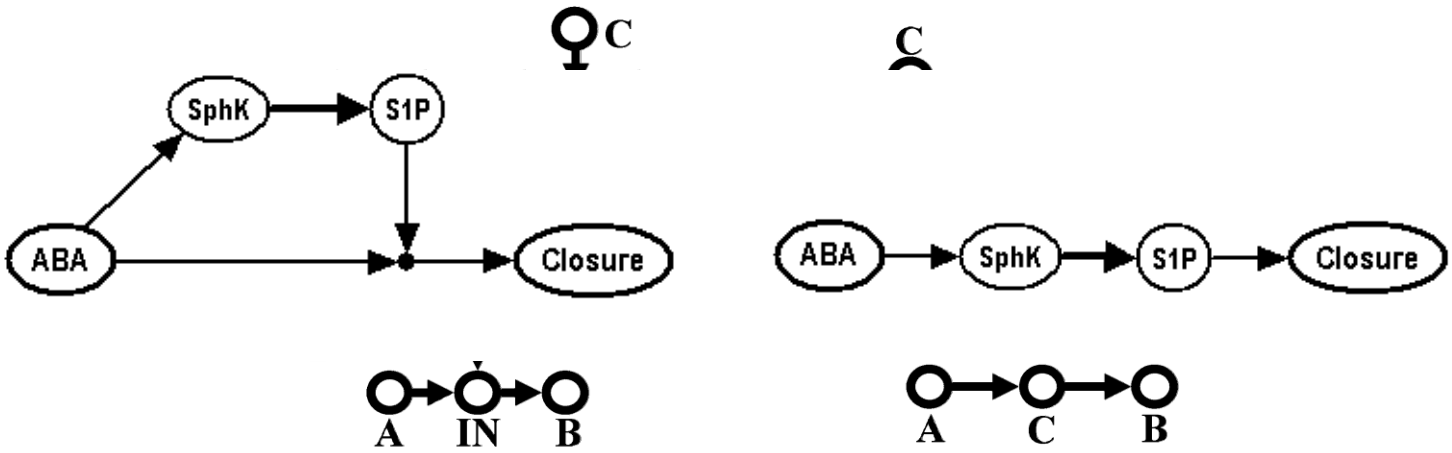
- activating or inhibiting effects on processes represented as intersection of two paths SphK ⟶⟶ (ABA ⟶⟶ closure)

Need to determine the closest regulator and target of each node

Network reduction

Find the most parsimonious (least redundant) network that incorporates all nodes and known processes.

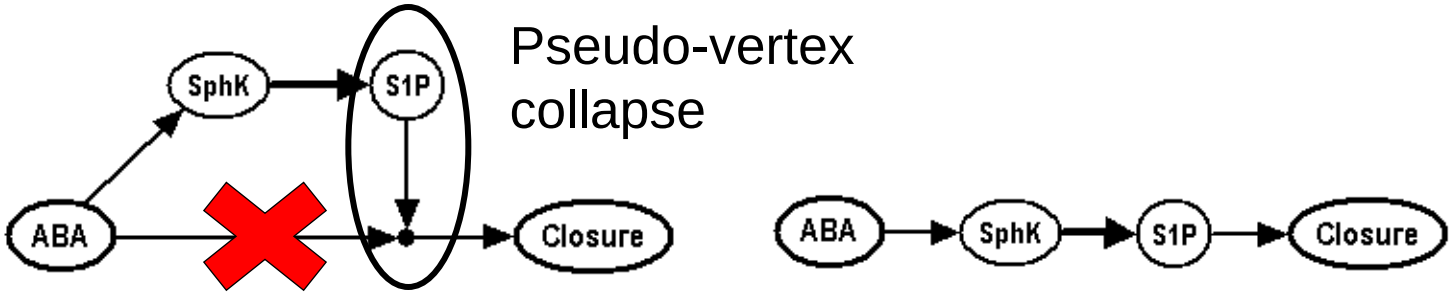
- Introduce intermediary nodes
- Contract intermediary nodes
- Review and revise



Network reduction

Find the most parsimonious (least redundant) network that incorporates all nodes and known processes.

- Introduce intermediary nodes
- Contract intermediary nodes
- Review and revise



Transitive reduction

General algorithm for network synthesis

Find the most parsimonious (least redundant) network that incorporates all nodes and known processes.

Two main algorithms:

1. binary transitive reduction with critical edges (BTR)

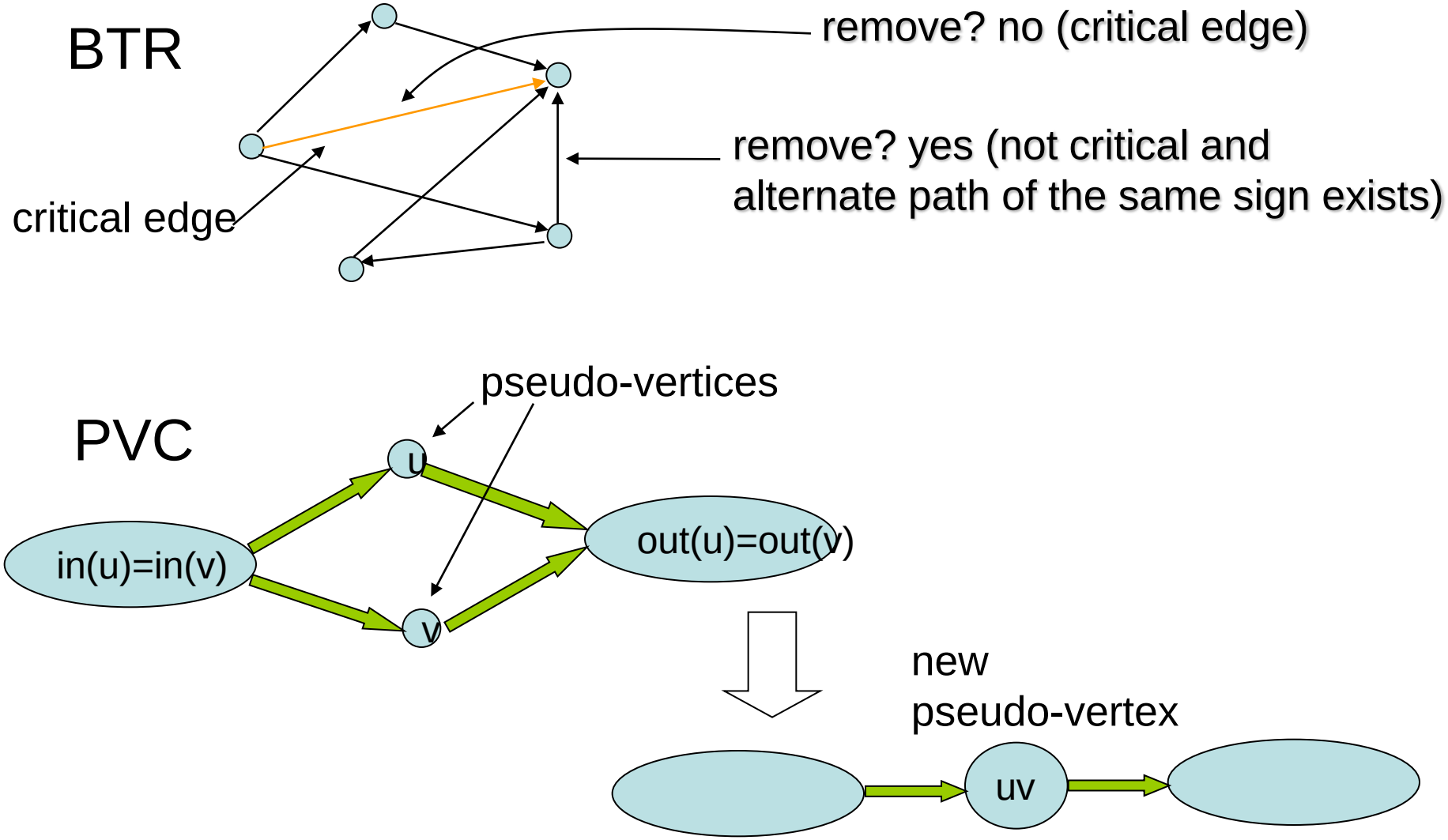
Transitive reduction: for each edge in the original graph there is a path in the reduced graph.
Binary transitive reduction: the sign of the paths needs to be maintained.
Critical edges: they correspond to direct interactions and should not be eliminated.

2. pseudo-vertex collapse (PVC)

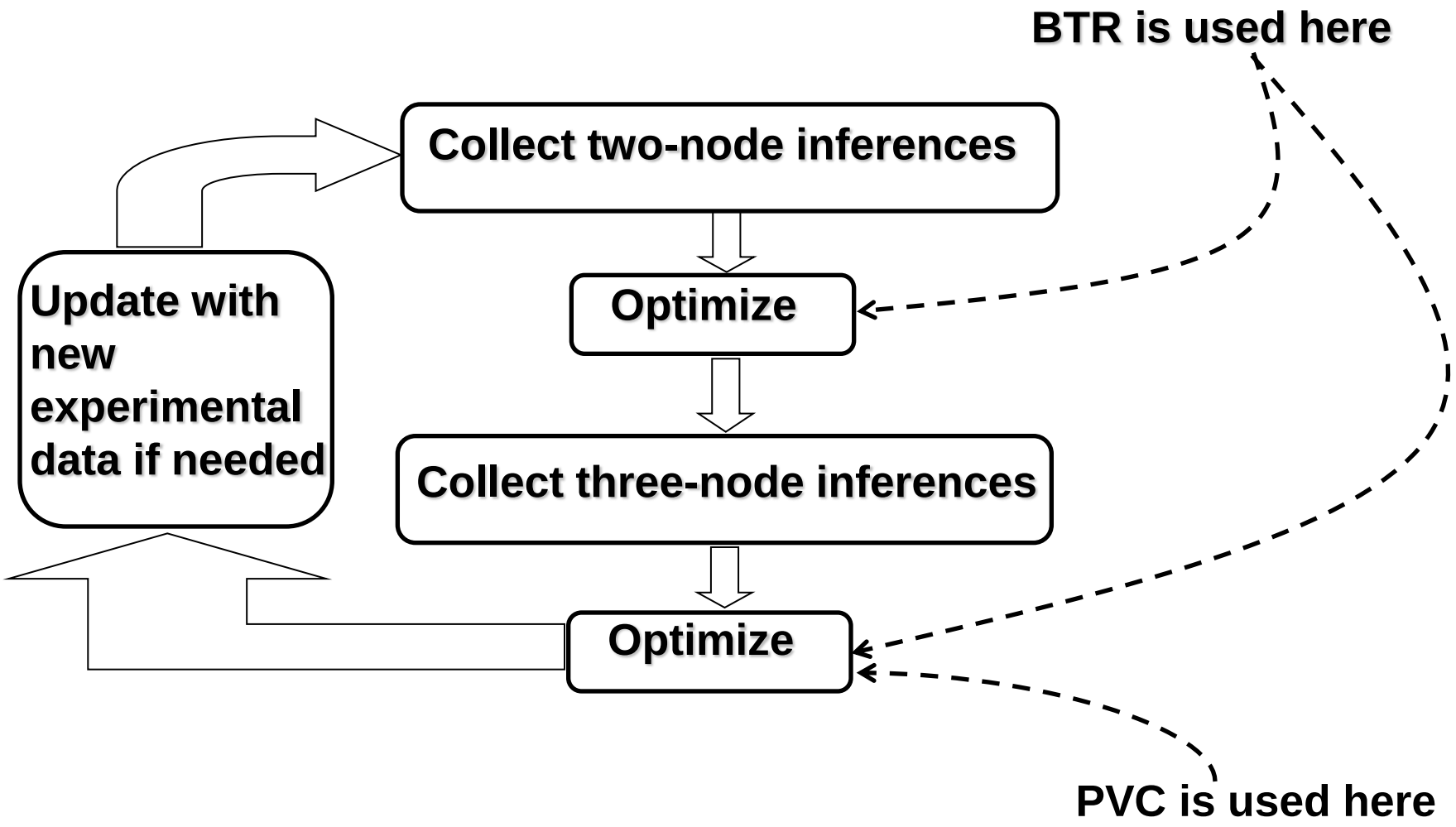
Merge a pseudo-vertex (intermediary node) with a real node or another pseudo-vertex if the nodes that can reach/can be reached from them, including the sign of the paths, are identical.

R. Albert, B. DasGupta et al, Journ. Comp Biology 14, 927 (2007).
S. Kachalo et al., Bioinformatics 24, 293 (2008).

An illustration of BTR and PVC



High level description of the entire network synthesis process



Implementation

NET-SYNTHESIS
http://www.cs.uic.edu/~dasgupta/network-synthesis/

Started from table of inferences collected by Li et al, compared the hand constructed ABA induced closure network with the one by NET-SYNTHESIS

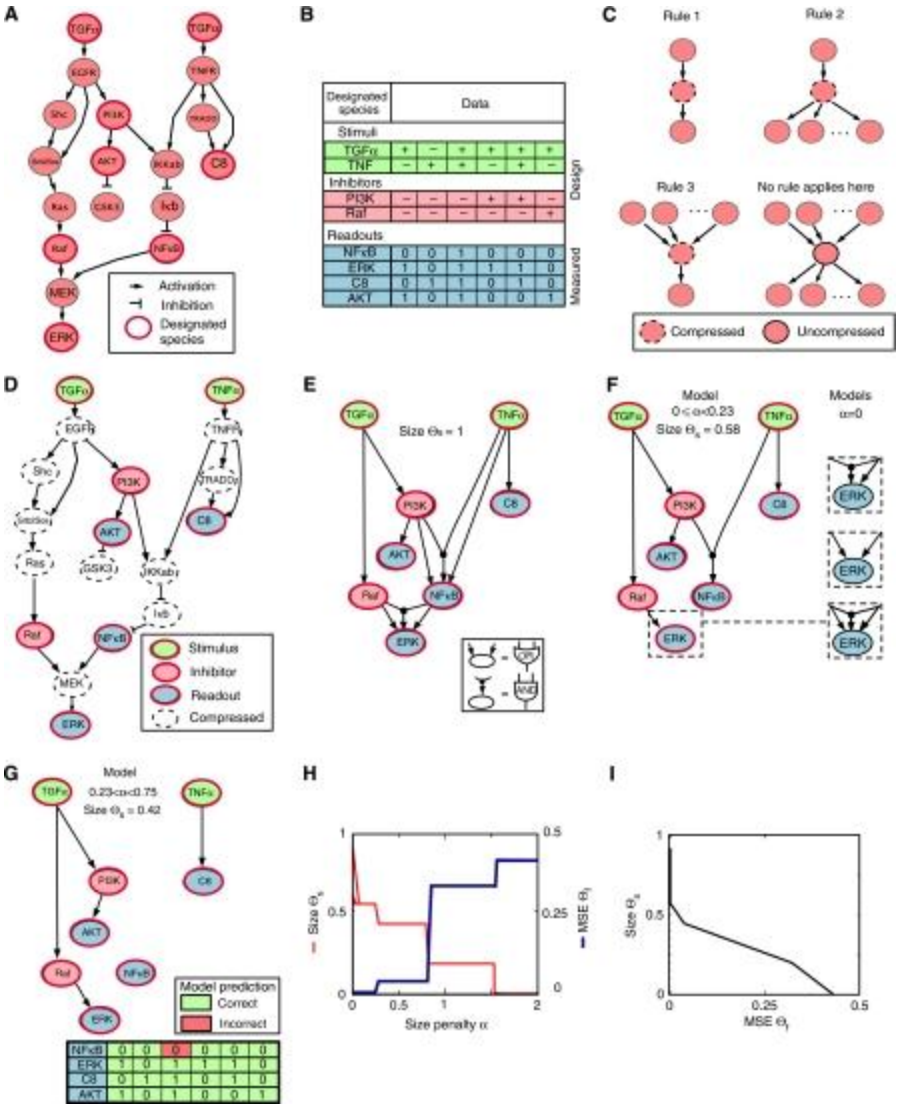
- Li et al. network has 54 nodes and 92 edges; NET-SYNTHESIS network has 57 nodes and 84 edges; 71 common edges
- Both networks have identical strongly connected component of vertices
- All the paths present in the Li et al. network are present in the NET-SYNTHESIS network as well
- The discrepancies are not due to algorithmic deficiencies but to human decisions.

- It took a few seconds to synthesize the NET-SYNTHESIS network
- Algorithm also useful for simplifying networks by designating nodes as pseudo-nodes and performing PVC.

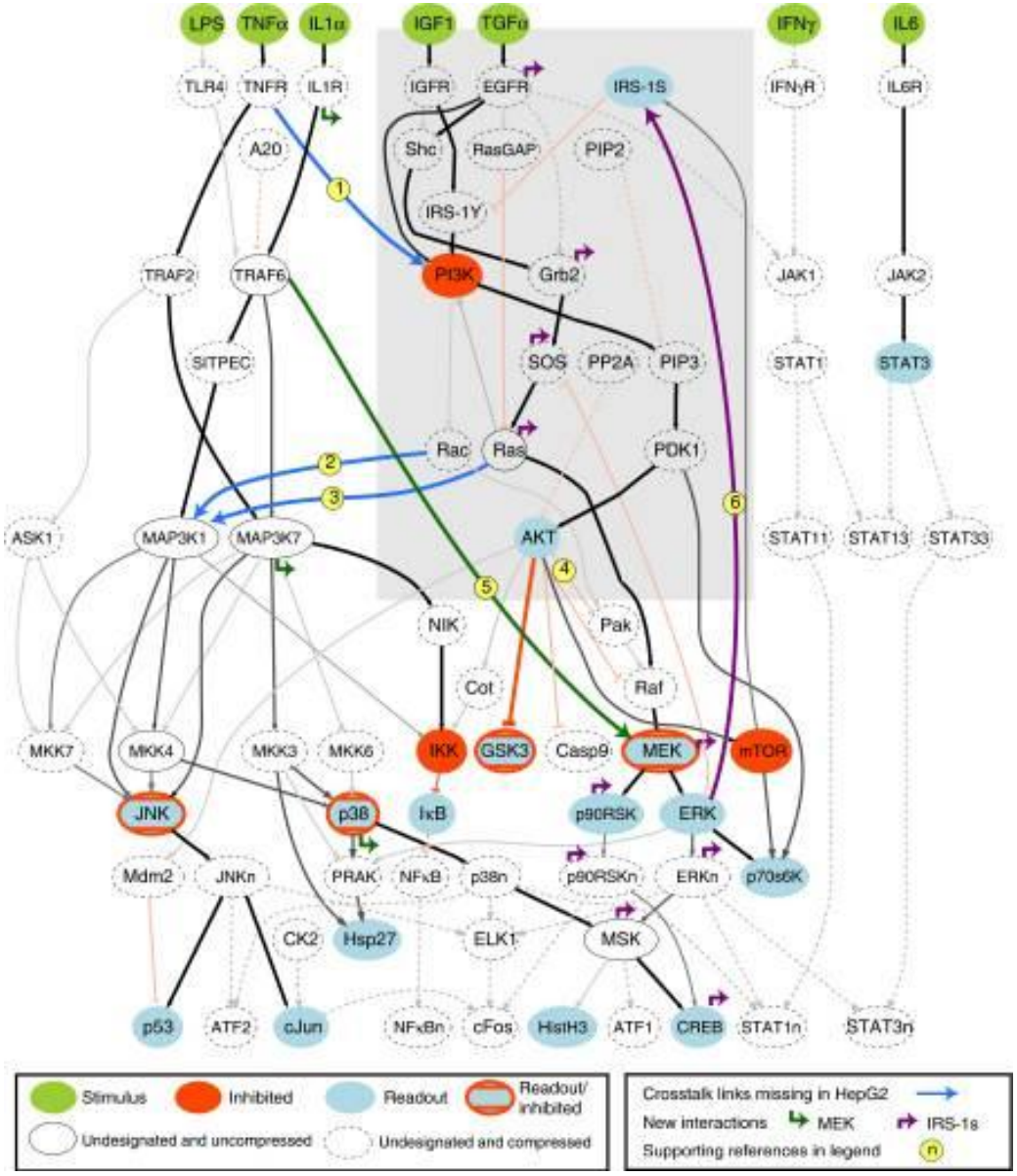
Inference of Boolean models from prior network and experimental data

- Construct an initial signal transduction network from the literature or databases
- Simplify the network
- Construct a meta-Boolean model that incorporates all possible Boolean rules
- Use experimentally obtained state information to find the simplest network/rule subset that explains the data
- Objective function for the fit trades off accuracy and sparseness
- Implementation: CellNetOptimizer

Saez-Rodrigues et al, Mol. Syst. Biol 2010



Inferred a new model of liver cancer cell signaling



- Original network: Ingenuity Pathways+ literature
- Found consensus of several solutions
- New experiments validate the model